Context and purpose of the challenge:
The objective of this challenge is to understand and predict the quality of a solution to a vehicle routing problem.

A quick reminder of the problem:
"Given a set of vehicles in a depot, and a set of customers and their requests, which vehicle must deliver which customer, and in what order, so as to minimize the total distance traveled by all the vehicles.
The goal is not to solve the problem itself, an algorithm has already been used to generate solutions! Thousands of solutions per instance! Indeed, this problem is a difficult problem and there is no perfect algorithm, which will give us the solution in polynomial time depending on the size of the instance (see the theory of NPcomplete problems). To obtain approximate solutions (sometimes optimal, but not guaranteed), we use metaheuristics, families of methods that will allow us to get good solutions. Here we are attacking a Capacitated Vehicle Routing Problem (CVRP) with 100 customers.

• *What is an instance?*
An instance is a configuration of the problem with a distribution of customers on a map, the location of the starting depot, etc. We have one file per instance, named for example XML100_2113_01.csv
• *Which algorithm was used?*
For each instance, a genetic algorithm variant (a metaheuristic) was applied for 10 min. This program thus generates several thousands or even tens of thousands (or more) of solutions per instance (per csv file).
• *What is the configuration of the problem?*
We have around 100 customers per instance, an unlimited number of vehicles available each with a certain capacity: each vehicle can satisfy 11 demand units. A customer seeks delivery of a product.
• *What is the objective to minimize?*
The algorithm seeks to satisfy customer demand while minimizing the distance traveled by the entire fleet of vehicles used: each vehicle used makes a round of customers and returns at the depot.
• *Purpose of the challenge?*
We are not looking at all to use machine learning to find good solutions (it would have been possible but here no), we will rather use the methods seen in class in order to be able to predict the quality of a solution without knowing its detail or its cost, but rather from certain characteristics related to the solution (see data description section).
• *Why such a goal?*
We believe that being able to predict and understand the determinants of a good or bad solution to a VRP problem could ultimately enable the implementation of hybrid methods combining machine learning and optimization. This would make it possible to take advantage of past experiences (the histories of solutions from which the FML algorithms learn) in order to guide, and in particular accelerate, the search for CVRP solutions thanks to optimization algorithms. This is a current research problem that we wanted you to benefit from… Because you are worth it!

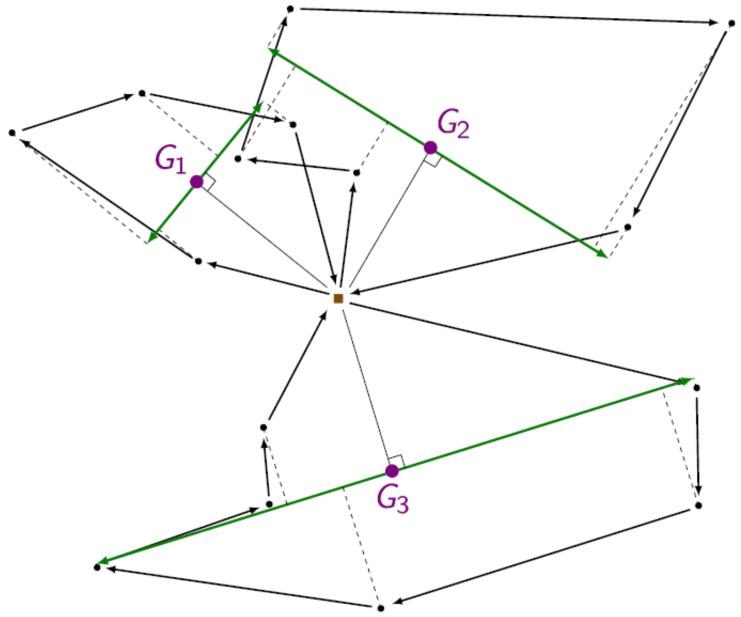
The file to download on the Moodle page of the course contains 4 sub-folders, named 2113, 2213, 3113 and 3213. Each of the 4 sub-folders corresponds to a sub-group of instances, classified (in the sense of clustering) according to the characteristics following:

2113: centered deposit, totally random customer position
2213: Centered repository, but clients form multiple clusters, geographic clusters.
3113: deposit near a corner of the map, random customer position
3213: depot close to corner, but clients form multiple clusters, clusters geographical.

In each subfolder, there are respectively 27, 27, 26 and 26 different instances, for a total of 106 instances. Each instance (each of the 106 files) contains thousands of lines. Each line corresponds to a solution: each solution/line was generated during the application of a genetic algorithm for 10 minutes.

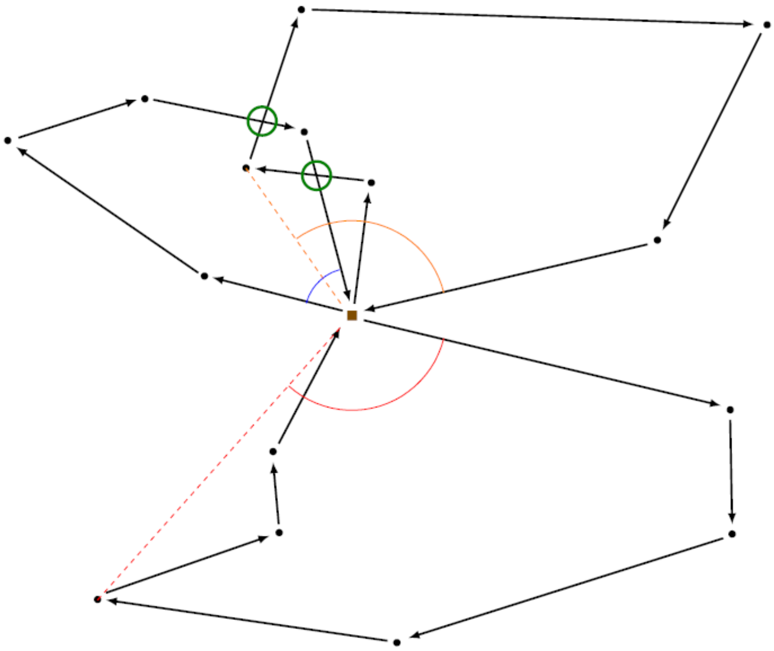
Explanation of data files:
1 line = 1 solution
Column 1: Instance name
Column 2: Cost of the solution
Column 3 to 20: the 18 characteristics/features of the solutions (which we name/rename S1...S18)
Note: Due to a bug, one of the features was not correctly calculated. It's up to you to detect and delete it.

S01 (column number 3) Average width of rounds This characteristic is equal to the average size of the green edges in the figure below. These correspond to the maximum distances of two customers on the same route, on the projection on the axis perpendicular to the depot-center of gravity axis.



S02 Standard deviation on the width of the routes

S03 Average size of rounds Characteristic equal to the average value of the maximum angles obtained between two customers of a route, and the deposit, represented below

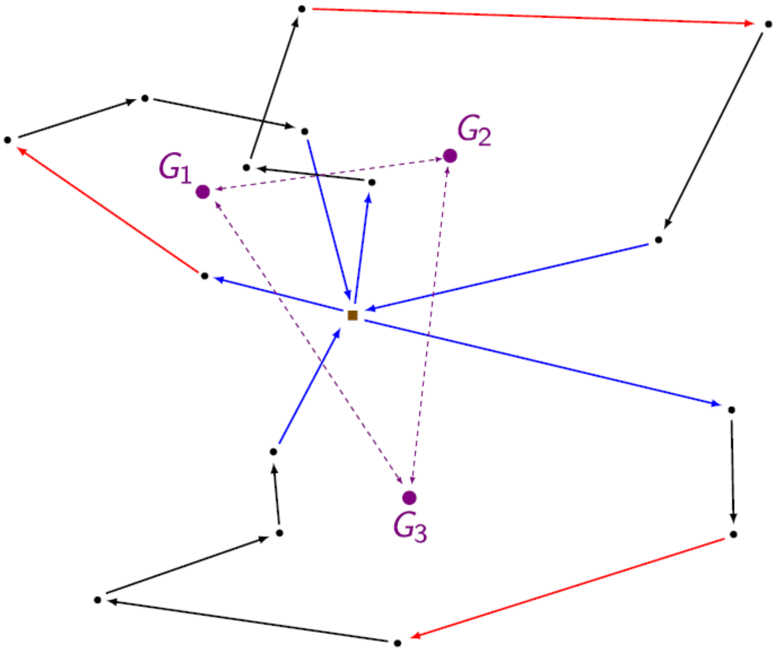


S04 Standard deviation on the span of rounds

S05 Average depth of rounds. Corresponds to the average distance per round between the customer furthest from the depot and the depot.

S06 Standard deviation on the depth of rounds S

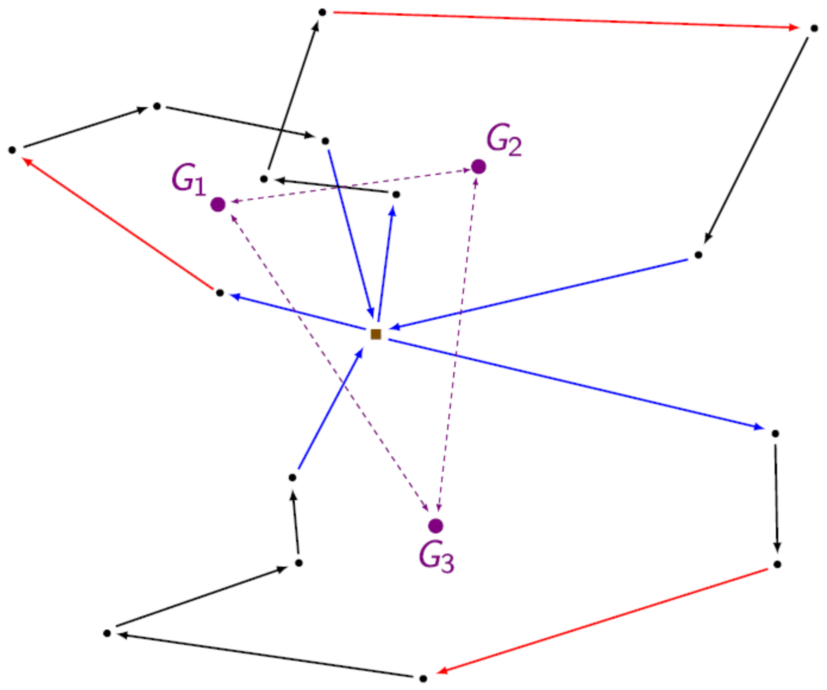
07 Length of first and last edge of each tour blue edges in the figure below, divided by the total length of the tour.



S08 Average length of the largest edge of each tour

S09 Length of the longest edge of each tour, divided by the length of the tour. This characteristic therefore indicates the proportion of travel time used for the longest "depot-to-customer", "customer-to-customer" or "customer-to-depot" trip of the route.

S10 Length of the largest interior edge of each route (edge not connected to the depot), divided by the length of the route. This characteristic therefore indicates the proportion of travel time used for the most long "customer to customer" movement of the route, and is represented by the red edges of the figure below.



S11 Average length of the first and last edge of each tour (blue edges in the previous figure).

S12 Demand of the first and last customer of each route, divided by the vehicle load. Indicates the average proportion of vehicle load due to the first and last customer.

S13 Demand of the customer furthest from the depot, for each round, divided by the vehicle load. Indicates the average proportion of the vehicle's load due to the remote customer
S14 Standard deviation on the request of the customer furthest from the depot

S15 Standard deviation along the length of each tour

S16 Average (Euclidean) distance between the rotated centers of gravity i.e: between the average coordinates of customer + depot, indicated by the points G1, G2 and G3 of the previous figure.

S17 Standard deviation on the number of customers for each round

S18 Degree of each mean neighborhood of customers. A customer who will be delivered after his nearest neighbor and before his 3rd nearest neighbor will have a neighborhood degree of 2.

Work to be done

The work consists of setting up a complete supervised learning methodology applying the methods seen in class. We propose to follow the following steps and indicate an indicative scale for each of them:

Descriptive statistics and feature engineering (2 points)

1. Uni- and multi-dimensional statistics: evaluation of data quality, understanding of the structure, links between variables
2. Recoding of variables, transformation, possible creation of new variables

Benchmark of regression methods to predict the cost of a solution to a CVRP problem (3 points)

The purpose of this part is to predict the "cost" variable, i.e. the second column of the set of data, thus formulating the problem as a regression. You will follow a classic supervised learning methodology (train/test), applying different methods seen in progress:

- Logistic regression
- kNN Regression
- Support Vector Regression
- Regression trees and random forest
- Gradient Boosting
- Neural networks and deep learning

The aim is to propose the best possible model in terms of generalization on a test set, i.e. to be able to predict the cost of a solution as precisely as possible, from the variables provided, or a subset of selected variables. The choice of the test set and the subtleties of evaluation will be important in our assessment.

Transformation into a classification problem (3 points)

This is the creative part of the challenge, which will decide between the best teams. For each instance (each seed file), it is obvious that the minimum cost value corresponds to the value of the best known solution. It is then possible to create a new column which can take the form of a binary variable (ex: good/bad solution) or categorical (e.g. bad/average/good/excellent). Threshold questions, which you you will encounter in your engineering career…. You then transform the regression problem into a classification problem and can apply the methods seen in class. You will propose a model capable of predicting the quality of a new solution and will evaluate it in generalization, on a test set. You can apply a class rebalancing method if necessary and calculate the usual performance metrics (confusion matrices, F1 score, etc.).

Quality of code and analysis (2 points)

Particular attention will be paid to the quality of your analysis, your ideas on the proposed problem. A very well commented notebook will be the minimum rendering